Regression Benchmarking with Simple Middleware Benchmarks

Lubomír Bulej, Tomáš Kalibera, Petr Tůma

DISTRIBUTED SYSTEMS RESEARCH GROUP
http://nenya.ms.mff.cuni.cz

CHARLES UNIVERSITY PRAGUE
Faculty of Mathematics and Physics
Benchmarking experience
- middleware comparison projects

Performance regressions
- scalability issues
- "creeping" degradation

Systematic benchmarking
- reveals performance regressions
- contributes to quality assurance process

Motivation
Regression Benchmarking

Variant of regression testing
- focused at performance
  - performance tracking
  - performance tuning

Differs from benchmarking in general
- tightly integrated with development
- fully automated
  - execution
  - measurement data acquisition
  - measurement data analysis
Existing Simple Benchmarks

Typical benchmark
- average duration of method invocation
- artificial workload
- simple results

Example
- note high relative variation
- hard to detect slight performance changes
High Variation Causes

Interference
- device interrupts
- context switches

Minimizing interference
- filtering \(\times\) not readily available
- robust estimators \(\checkmark\) median
- shorter intervals \(\checkmark\) low-level operations
Robust Estimators

Median …

- not affected by small number of exceptional observations

Getting enough data

- enough for accurate analysis, but not more
- order statistics or other precision estimates
Shorter Intervals

Marshaling …
- less interference
- results better traceable
to code

Example
- note still high relative
 variation
Impact on Resolution

Duration $D$

$D_1, D_2, \ldots, D_N$

$$s_D = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (D_i - \bar{D})^2}$$

$$V_D = \frac{s_D}{\bar{D}}$$

$$\bar{D} = \frac{1}{N} \sum_{i=1}^{N} D_i$$

$k \cdot d_1, k \cdot d_2, \ldots, k \cdot d_N$

$$s_d = \frac{s_D}{\sqrt{k}}$$

$$V_d = \frac{s_d}{d} = \sqrt{k} \cdot V_D$$

$$\bar{d} = \frac{\bar{D}}{k}$$
Do Not Hope for Zero Variation

Application state
- heap, JIT …
- mitigated by warm-up

System state
- physical memory mapping …
- persists thru warm-up
- more visible when measuring shorter intervals
Comparing Results of Consecutive Runs

What we assume about median estimate

- difficult to make any assumptions
- independent identically distributed
- asymptotically normal distribution

Standard statistical tests

- two-sample variance f-test to validate equal variance assumption
- unmatched two-sample t-test to compare averages of a pair of benchmark result sets
Real-World Data

TAO weekly builds
- from May 19, 2003 to Oct 27, 2003
Conclusion

Middleware regression benchmarking
- variant of regression testing
- differs from benchmarking in general

Simple middleware benchmarks
- easy way to track and detect changes in performance
- artificial workload, complex benchmarks needed to exercise the interplay of various subsystems
Future Work

Complex middleware benchmarks

- realistic workload tests a set of features
- issue is the analysis of benchmark results